DOCUMENT RESUME

ED 094 006 95 TM 003 864

AUTHOR O'Connor, Patricia; And Others
TITLE Improving Reliability in Assessment of Technic Products.
SPONS AGENCY National Institutes of Health (DHEW), Bethesda, Md. Bureau of Health Manpower Education.
PUB DATE [Apr 74]
NOTE 9p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Dental Schools; *Evaluation Criteria; *Performance Tests; Rating Scales; *Student Evaluation; *Test Reliability

ABSTRACT

For instructional materials to be certified as "effective," students must meet instructional objectives operationalized by criterion tests. By implication, evaluators must agree when criteria are or are not met. Fourteen instructors evaluated 10 posterior bridges. Interjudge agreements for total bridges and individual attributes were low, as they tend to be whenever dental technic products are evaluated. A method for developing more reliable rating forms is described. It consists of: (1) limiting discriminations to the dichotomous decision "acceptable", "unacceptable"; (2) initially resolving differences among faculty; and (3) defining characteristics of acceptability on observable terms, and providing a photographic example of a minimally acceptable product for each attribute. (Author)

ED 094006 1.

TM 003 864

Improving Reliability in Assessment of Technic Products

Patricia O'Connor, Robert E. Lorey and Arun Garg
The University of Michigan School of Dentistry

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Studies indicate that agreement among teachers in assessing the quality of student technic products is low. Furthermore, the evidence for improving consistency among raters through training programs has been discouraging. Low agreement among faculty raters is found not only in assessments of interjudge reliability, but instructors also differ widely in severity of grading. Lack of agreement creates serious problems for students. Equity is compromised since students' grades are influenced by who evaluates their technic products. Inconsistent standards reduce the value of diagnostic feedback. Questionnaire data indicate that students frequently identify the low level of faculty agreement as an instructor behavior that interferes most with their learning. For example, one instructor may describe a cavity preparation as "too deep," a second as "too shallow," and a third as "acceptable." Stories which may or may not be apocryphal report students, whose work is criticized at a checkpoint taking the unaltered product to another instructor, or even to the same instructor, to be told their work is acceptable or even excellent. One investigator has drawn a parallel between evaluations in technic courses and studies in which "experimental neurosis" is produced in laboratory animals by presenting problems in which the correct response is deliberately varied over trials.

The work reported here was supported in part by Grant # OEG-0-8-1220-01-02 of the Department of Health Education and Welfare, Bureau of Health Manpower Education. AERA 2806.

The problem becomes critical when one attempts to develop and evaluate instructional materials designed to teach students to make products which meet performance criteria. In the absence of clear performance standards, the certification of materials as "effective" is not possible. Our work indicates that one requisite to producing acceptable products is learning to discriminate between those that meet and fail to meet specific criteria. Observations in the laboratories have shown that in a substantial proportion of student-faculty interactions students ask faculty whether or not a particular step in creating a product is satisfactory. Moreover, when students were asked in a questionnaire if presenting slides showing ideal and flawed preparations before they were required to make these preparations themselves would contribute to their learning, over eighty percent responded affirmatively. Our initial attempts to provide this instruction have received a positive student response. Instructional units to teach the attributes of a good product are required. To teach students the standards for which they should strive and by which they may evaluate their work necessitates that faculty first agree upon standards of evaluation.

This paper reports one study in a continuing effort to develop methods which will result in consistent evaluation among faculty judging technic products. The product to be judged was the preparation of a tooth for a posterior bridge. The rating form normally used in evaluating the preparation is comprised of six criteria each of which is to be applied on a five point scale; clinically unacceptable, minimally acceptable, adequate, good, and excellent. A rating of "adequate" implies that the student has met course objectives for the criteria applied. On the rating scale only the attributes

themselves and the rating scale are listed. The purpose of this study was to test whether providing for reference six color slides each of which illustrated "adequate" meeting of one criterion would result in improved faculty agreement. Agreement as used here refers not only to interjudge reliability but also to reduced variability among instructors in stringency or leniency in rating.

Method

Ten faculty members who instruct students in the sophomore course in preclinical crown and bridge were subjects in the study. The names of 40 students were selected at random from the class list of 144. Twenty were assigned at random for evaluation in the first rating session and the remainder for evaluation in the second rating session. Preparations were assigned numbers to disguise the identity of students. Ratings were made individually by instructors under normal illumination with added light from a tensor lamp. Instructions to faculty for the first rating session were as follows:

> The following items are criteria for evaluating the quality of the molar preparation for sophomore students' bridges. Please evaluate each attribute in terms of the five point scale provided. In making your judgments, please take the orientation that ratings of 3, 4, and 5 indicate that on the attribute rated the student has met the course objective. Ratings of 1 and 2 indicate that the student has fallen far short or somewhat short of meeting the course objective.
>
> Please use all parts of the scale and remember that a rating of 3 or better implies that for the attribute judged the student has met the course objective. A rating of 2 or 1 implies the student has failed to meet a course objective.

Ratings were made and recorded. From these ratings we selected six preparations each of which would be used to illustrate "adequate" meeting of one of the six criteria. Faculty consensus was the basis

of selection. Ideally, each selected preparation would have been rated "3" or "adequate" by all raters. Since complete accord was achieved in no instance, we selected for each criterion that preparation for which the mean rating was closest to "3" and for which variability was minimal. One of the authors (REL) who is a senior member of the Crown and Bridge Department photographed each of the selected preparations from the vantage point at which the particular characteristic could be best viewed. From the resulting color slides in which preparations were equal in size to the preparations themselves, he and a second member of the Crown and Bridge Department selected the one which most clearly illustrated adequate attainment for each of the six criteria.

In the second evaluation session, the six slides were presented simultaneously in a lighted view box and were labelled with the attribute illustrated. Instructions to faculty were identical to those in the initial rating session with the following addition:

> The illustrations provided represent faculty consensus of a rating of "3", i.e. "adequate" attainment for sophomore performance objectives for the attributes illustrated. In making your ratings, please assume that the example provided warrants a rating of 3. Use the illustrations for reference in making ratings of all preparations.
>
> PLEASE CONSIDER ONLY THE INDICATED ATTRIBUTE FOR A GIVEN PREPARATION. The process to follow in each case is:
> Assuming that the photograph illustrates "adequate" attainment for the attribute in question, what rating should be assigned to the preparation rated?
> - If it is substantially inferior to the preparation shown for that attribute, assign a "1".
> - If it is somewhat inferior to the preparation shown for that attribute, assign a "2".
> - If it is equal in quality to the preparation shown for that attribute, assign a "3".
> - If it is somewhat superior to the preparation shown for that attribute, assign a "4".
> - If it is substantially superior in quality to the preparation shown for that attribute, assign a "5".

Results

In the results, faculty agreement in the initial session and the session with slides are compared, first with regard to inter-judge reliability and second, with regard to stringency of standards.

To assess interjudge reliability for the preparations as a whole, ratings for each instructor were summed over all attributes for each preparation. The possible range of scores for a given preparation is from six to thirty. Product moment correlations between each judge's scores for the twenty preparations and the combined scores for the remaining nine judges were computed for the first and second rating sessions. The mean r for the initial rating session was .70 and for the second .83, showing a mean increase of .13. Of the ten instructors, r was higher in the second session for eight, the same for one and lower for one. This result was significant at the .01 level using the Wilcoxon matched-pairs signed ranks test.

Since we were interested not only in reliability of assessment as a whole but also in reliability for each of the six criteria, we examined faculty agreement for each of the attributes separately. For this purpose we employed the coefficient k developed by Jacob Cohen. Ratings were collapsed to three categories; unacceptable (ratings of 1 and 2), acceptable (rating of 3), and superior (ratings of 4 and 5). For each pair of raters, scores were cast in three by three tables. Agreements consist of those instances in which both raters have assigned the same ratings for the twenty preparations judged on a given criterion. The coefficient k is simply the proportion of agreements that occur after chance agreement is removed from consideration. In the analysis k's were

computed for all pairs of raters in the initial session and in the slide-present rating sessions. For each pair of raters the direction of differences in k between the two rating sessions was recorded. Higher positive k's in the second session would indicate higher reliability on individual attributes. Results using the sign test indicated that for three attributes interjudge reliability was higher in the second rating session; p values were $< .02$, $< .002$ and $< .001$ for these attributes. For the remaining three attributes no improvement was shown, nor did trends even approach statistical significance. The markedly improved reliability for three variables and its absence on the remaining three is difficult to attribute to non-systematic factors. Our post hoc explanations bear on the inadequacies of the slides for the particular attributes shown. A discussion of our hunches on this point would however require describing more than you would care to know about molar chamfer preparations.

We wished also to investigate whether, independent of changes in interjudge reliability, providing slides to define "adequacy" on each attribute reduced the differences in stringency or leniency in grading among instructors. We selected a method of analysis for this question based upon some reasoning I would like to describe briefly. If raters differ consistently among themselves in stringency of grading it follows that on each of the twenty preparations, the magnitude of the summed numerical score will systematically vary with the instructor. In the most extreme case the same instructor would consistently assign the highest score, another consistently assign the second highest score, another the third highest score and so forth. In this extreme case, rho's or rank order correlations computed between scores on any pair of preparations would result in

rho's of 1. To the degree that instructors did not vary in the stringency of standards applied, the rho's would approach 0.

To test the effects of the experimental treatment on instructor bias Spearman rho's were computed between all pairs of preparations. In this analysis each instructor is a subject and his score on each preparation, the equivalent of a score on some variable or attribute. For example, the scores might be analagous to twenty tests in arithmetic or spelling. Rho's were computed for the first rating session and compared with those computed for the second. The results of the analysis showed that the median rho for the first rating session was .52 and for the second .21. When rho's were broken at the combined median for both sessions, 73% of those for the first and 28% of those for the second fell in the "high" group. The chi square resulting from this analysis is 77.86 which is significant beyond any level recorded in statistical tables. Although the inflated N makes chi square analysis inapplicable, the obtained differences seem to imply that the stringency differences among instructors have been sharply reduced.

Changes in instructor stringency bias between the first and second session may be discussed descriptively. The standard deviation for instructor mean ratings in the first session was 2.36 and in the second 1.28. Similarly the average deviations were 1.72 and .88 respectively. The F ratio could not be used to test the difference in variability between the two rating sessions because of departures of scores from normality. In the first session, in fact, the deviations of instructors' mean scores from the combined mean were bimodally distributed. For four instructors the deviation was less than half a point and for two others about a point. However four instructors

showed extreme bias, i.e., two instructors assigned mean scores more than two and one half points above the combined mean and two others, mean scores more than two and one half points below the combined mean.

In the rating session with slides, the mean ratings of three of the four deviant instructors were reduced and were within one point of the combined mean for that session. The most severe grader's scores were half a point closer to the combined mean, but his mean score was nevertheless over three points below the combined mean. The remaining six raters as a group showed no evidence for change. In the initial session, the mean deviation score for them was .31 and in the second .44. For these subjects as a whole, the absence of change is not a serious matter of concern since, in the first session, they had not shown stringency bias. The dramatic change shown in the rho analysis is then attributable to the sharp drop in instructor bias for the three of four instructors whose grades changed markedly toward the combined group mean.

After the second session, in a questionnaire item, instructors were asked to predict whether their agreement with the combined faculty rating would be higher in the first or second session and to state the reason for their prediction. Eight of the ten predicted improvement on the second round and two were uncertain. All instructors who predicted improvement referred to the positive contribution of common standards for establishing a norm.

Discussion

The results of the study support the positive contribution of slides illustrating the minimum requirement for meeting a criterion in a technic preparation to instructor agreement. They encourage

the continuation of efforts to develop instructional materials that will teach students to learn criteria for evaluation of a product. Faculty themselves will of necessity be deeply involved in specifying criteria and selecting models and photographs. It is our hope that the materials developed will be used not only for student instruction but for faculty instruction as well. If we are successful in developing materials for this purpose, we will be able to evaluate the effectiveness of instructional materials designed to teach students to produce technic preparations that satisfy standards of excellence.